EXCLI Journal 2005;4:87-90 – ISSN 1611-2156
Received: 26. September 2005, accepted 5. October 2005, published: 7. October 2005

Review:

The necessity of extraction of ligand binding data from literature

Sayed-Amir Marashi

Department of Biotechnology, Faculty of Science, University of Tehran, Enghelab Ave., Tehran, Iran, Fax: +98 21 66491622, E-mail: marashie@khayam.ut.ac.ir

ABSTRACT

After identification and validation of targets in a drug development process, some independent lead compounds should be selected and optimized for their activities. Then, safety assessments and clinical trials decide whether the drug is proper to enter the market. Different stages of this process (and especially the identification of lead compounds) are extremely expensive and time-consuming. Rational drug development methods try to reduce the costs by optimizing the pace of drug discovery and reducing the number of products abandoned during development. For decades, many investigators have studied the ligand-protein interactions, but very few structured databases are devoted to such information. Herein, development of such databases is proposed, since it is obvious that our prior knowledge about the chemico-biological interactions can help us choosing appropriate lead compounds without further experimental and computational investigations, which are usually based on searching in gigantic combinatorial databases of chemical compounds.

Keywords: Mining literature, ligand binding, drug discovery, lead compounds

Biological activities are almost always associated with the interaction of two or more biomolecules: ligands bind to carrier proteins in blood circulation; substrates bind to enzymes before the reaction takes place; bindings of hormones to receptors trigger signals inside cells. Because of the importance of the subject, binding of ligands to proteins and other biomolecules (like DNA) has been studied extensively for decades.

Many drugs perform their activity via interference with the normal action of certain compounds, typically with an inhibitory mechanism. Therefore, study of the biochemical properties of ligand-protein (or in some cases, ligand-DNA) interaction has a central role in the discovery of new drugs, and is of commercial importance.

Drug discovery entered a new era when results of the Human Genome Project were released (Lander et al., 2001). It is now possible to search proteins of differential expression patterns in the transcriptome and/or proteome of a diseased tissue. This target identification and validation is normally the first step in the drug development process (Figure 1). In the next step, proper inhibitors (or activators) for the related activity should be found (Nuttall, 2001), and their activity should be optimized by adding, removing and changing the compound substituents. Safety and toxicology assessments and clinical trials are the last steps before a drug can go to the market. It is obvious that our prior information

about the nature of these inhibitions (or activations) is particularly important, and speeds up the drug discovery process, and effectively reduces the development costs.

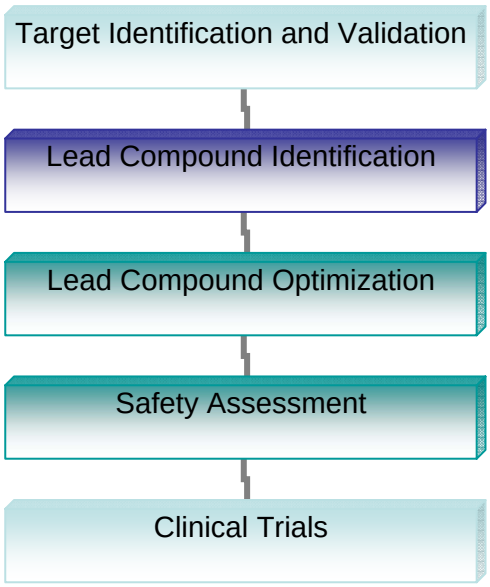

Figure 1: Drug development process.

Although SAR and QSAR surveys are traditionally used in rational drug discovery, existing *in silico* drug design techniques also robustly rely on the study of three-dimensional structure of proteins associated with other molecules (Joseph-McCarthy, 1999). Different modes of ligand-protein interactions have been studied in details using such data. At present, many researchers try to perform the task of "docking" novel ligands on the surface of a receptor or an enzyme of known structure, in order to find possible binding sites. This method seems promising, but accurate high-throughput screening for a ligand with desired properties can be extremely time-consuming, since the bigger the combinatorial databases of chemical compounds, the higher the probability of finding a better potential drug. One possible modification of this approach is to restrict our search dataset to analogs of ligands of known binding affinity towards the target, as the probability of finding a ligand with desired properties in a dataset of analogs (structurally related molecules) is greater than such probability in a huge dataset of (random) molecules. It is obvious that our prior knowledge of ligands with known properties again helps in solving our problem. Moreover, the need for knowing the three-dimensional structure of target protein is eliminated in this approach. In fact, design of lead compounds on the basis of known ligands of a target protein is a major strategy in the development of new drugs (Twyman, 2004).

As a result of large amounts of published information on molecular binding properties, it seems reasonable to register these data in specific databases. Today, one can find a wide variety of databases with different scopes, such as genomic information, protein structures, and cellular expression levels of biomolecules. Surprisingly, there are only a limited number of databases with the subject of ligand binding. Most of these databases collect information about ligand binding properties of complexes

with known three-dimensional structures (for example see Chen et al.,2002a; Wang et al., 2004; Golovin et al., 2005). BindingDB (see Chen et al., 2002b; 2002c; Available from: http://www.bindingdb.org) is the only database completely devoted to the experimental results of the interaction of different compounds, typically with low molecular weights, with biological macromolecules like proteins. At the moment, this database contains only experimental data obtained from isothermal titration calorimetry (ITC) and enzyme inhibition assays. At the moment, the database contains only a few thousand ITC and enzyme inhibition entries. Clearly this is only the tip of the iceberg: a massive amount of information on ligand binding (obtained from experiments that take advantage of a variety of techniques) has been accumulated in literature for years, and it is still growing in size.

Mining data from literature is not an easy mission. It is practically impossible to study manually all the previous publications (or even their indexed abstracts) in the fields of biochemistry, biophysics, molecular biology, pharmacology and biomedical sciences; for example, there are at least twelve million abstracts indexed in Medline (http://www.ncbi.nlm.nih.gov/PubMed/). Simple searches in literature databases using definite keywords (as well as making use of MeSH vocabulary) will not work as well, because of the large number of false positive and/or false negative hits. As a result, it seems that the most reasonable method must be automatic capture of information using special computer algorithms. At present, no special algorithm is used in BindingDB to extract data from literature (Y. Lin, personal communication).

Automatic literature searching for mining data has been reported previously to extract important biochemical and biomedical information (Shatkay and Feldman, 2003; Cohen and Hersh, 2005; Krallinger et al., 2005). Such methodologies can be applied to the subject of ligand-protein interactions, in order to highlight suspected papers; there will be a need to browse the related articles by hand in the next step, to make sure about the content and then to extract the data lied within the body of the texts and tables. Without such automatic approaches, construction of an integrating ligand binding database seems to be extremely intricate. At present, in our group we deal with such a problem, and we hope to develop a simple and effective algorithm to mine such data from literature.

References

Chen X, Ji ZL, Zhi DG, Chen YZ, CLiBE: a database of computed ligand binding energy for ligand-receptor complexes, Comput Chem 2002a; 26: 661–666.

Chen X, Lin Y, Gilson MK, The Binding Database: Overview and User's Guide, Biopolymers 2002b; 61: 127–141.

Chen X, Lin Y, Liu M, Gilson MK, The Binding Database: data management and interface design, Bioinformatics 2002c; 18: 130–139.

Cohen AM, Hersh WR. A survey of current work in biomedical text mining. Brief Bioinform 2005; 6: 57–71.

Golovin A, Dimitropoulos D, Oldfield T, Rachedi A, Henrick K, MSDsite: A Database Search and Retrieval System for the Analysis and Viewing of Bound Ligands and Active Sites, Proteins 2005; 58: 190–199.

Joseph-McCarthy D, Computational approaches to structure-based ligand design, Pharmacol Ther 1999; 84: 179–191.

Krallinger M, Erhardt RAA, Valencia A. Text-mining approaches in molecular biology and biomedicine. Drug Discov

Today 2005; 10:439–445.

Lander ES, Waterson RH, Collins FS, Initial sequencing and analysis of the human genome, Nature 2001; 409: 860–921.

Nuttall ME, Drug discovery and target validation, Cells Tissues Organs 2001; 169: 265–271.

Shatkay H, Feldman R. Mining the biomedical literature in the genomic era: an overview. J Comput Biol 2003; 10: 821–855.

Twyman RM, Principles of proteomics. BIOS Scientific Publishers, Milton Park, Abingdon, Oxon, England, 2004. (see pp. 229–230).

Wang R, Fang X, Lu Y, Wang S, The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known three-dimensional structures, J Med Chem 2004; 47: 2977–2980.